TECHNICAL UNIVERSITY OF DENMARK

5 AUGUST 2024

THESIS DEFENSE

Unveiling Diversity in Wikipedia: Analysis of Human Dimensions Across Pages and Contributors

Alejandra Navarro Castillo

Human-centered artificial intelligence

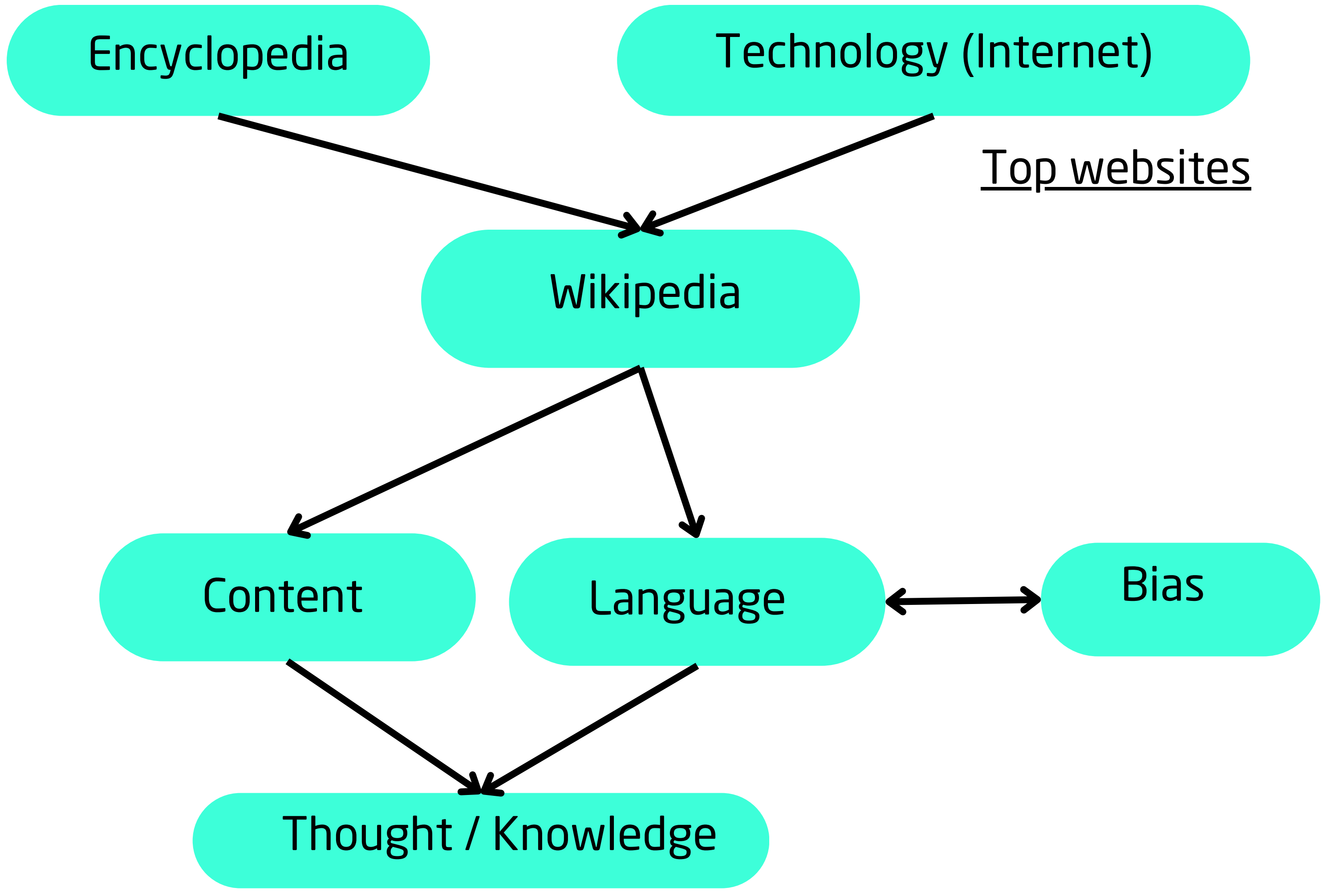
OVERVIEW

- Motivation
- Scope: objectives
- Review of Related Literature
- Dataset
- Methods and results: Who
- Methods and results: What
- Methods and results: How
- Summary and Conclusion
- Future research
- References
- Q&A

Link to my repository: click here

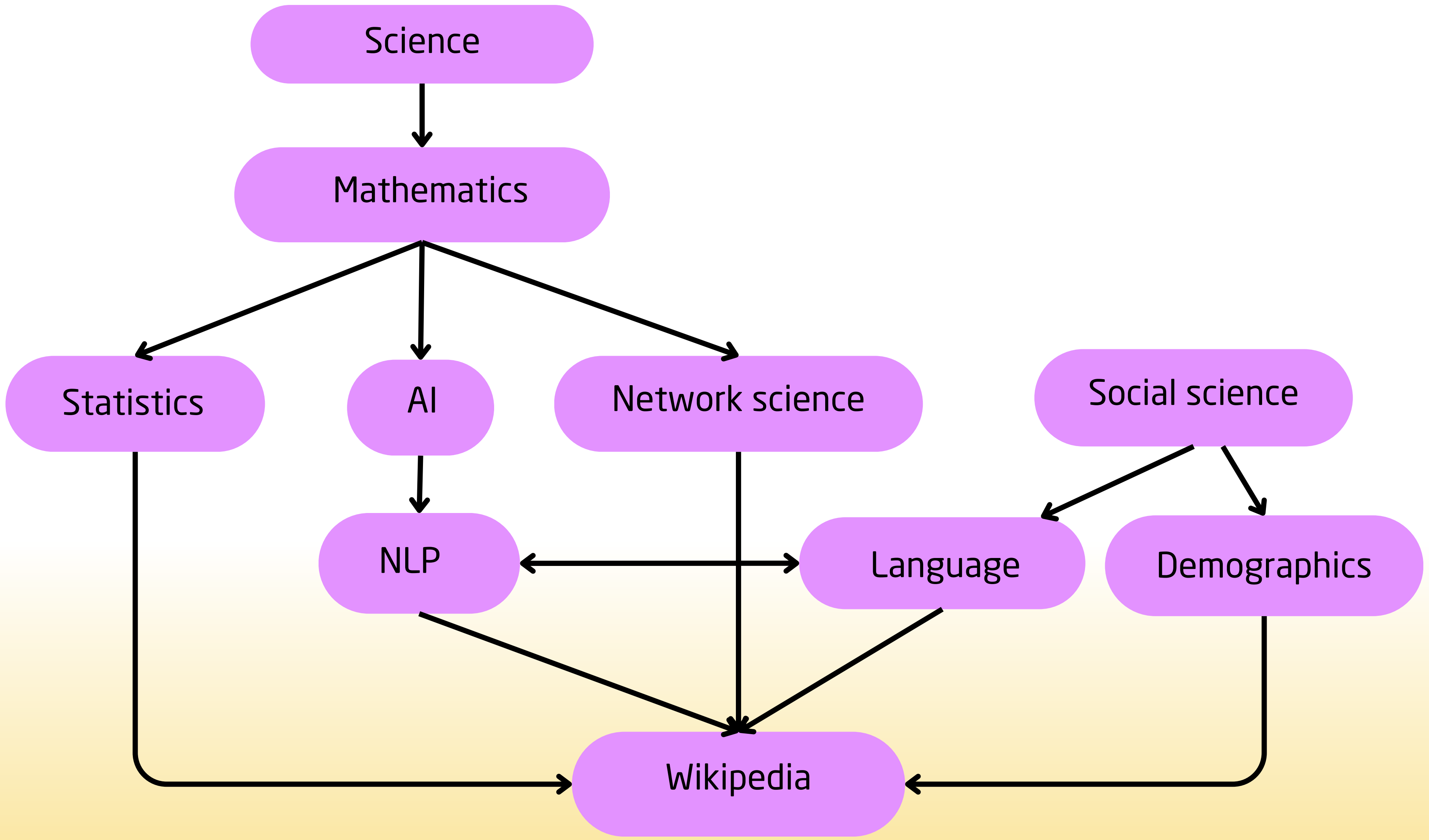
MOTIVATION

Our motivation



Top websites

My motivation



BACK TO OVERVIEW →

SCOPE: OBJECTIVES

WHO? WHAT?

- Gender distribution among Wikipedia editors?
- Gender distribution among Wikipedia biography pages?

HOW?

- How does gender influence the text content in biographies in Wikipedia?
- Are there any differences in gender in the hyperlinks network of articles?

Editors and biographies gender distribution

Text content

Hyperlinks structure

BACK TO OVERVIEW →

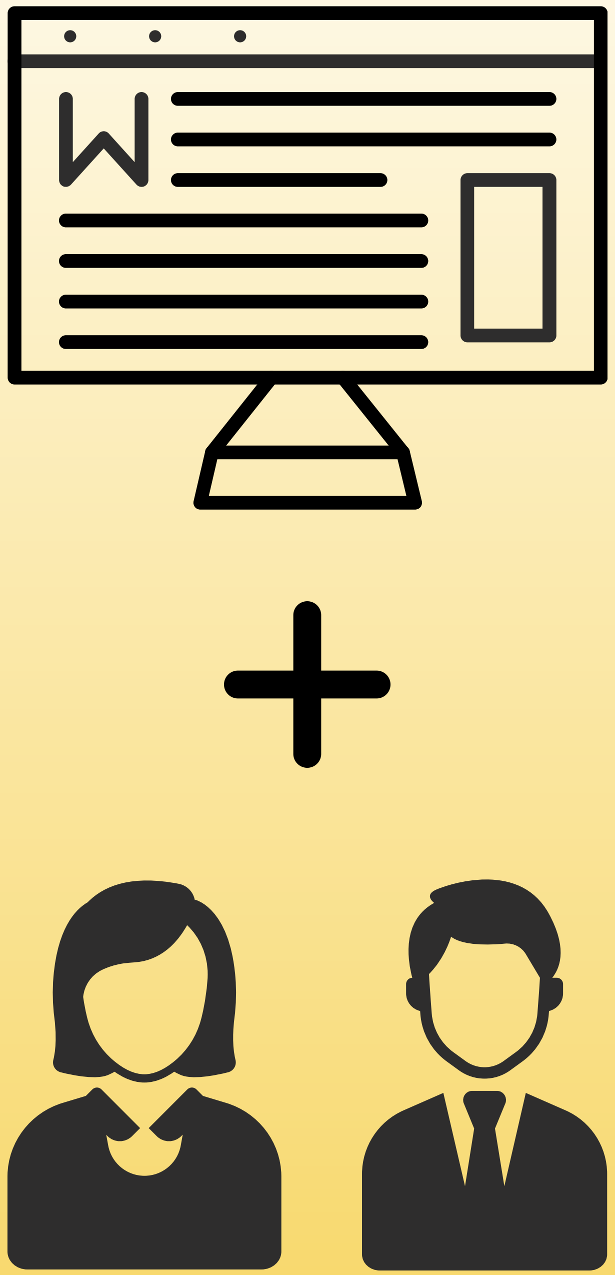
REVIEW OF RELATED LITERATURE

Previous studies

- They explore how gender is depicted in Wikipedia by textual and visual elements. [BW22; GLM15; Wag+15; Wagner+16; Bey+22; ZFW17; YWK16]
  - Length of articles
  - Lexical analysis
  - Visual elements
  - References

My work

- It adds an extra layer of analysis by also differentiating between the gender of those who create and edit these biographies.



BACK TO OVERVIEW →

MY DATASET

Initial dataset

- Wikipedia dump with revisions from 2001 to 2023 of English Wikipedia.
- Editors: Excluded anonymous users and bots.
- Pages: filtered for only biographies with gender information.

What I carried out:

- Pages: filtered for only last revision of each page.
- Editors: Group editors for every revision of a page.
- Pages: retrieve Wikitext and Wikilinks through Wikipedia API.
- Pages: clean Wikitext.

BACK TO OVERVIEW →

Methods and results: Who

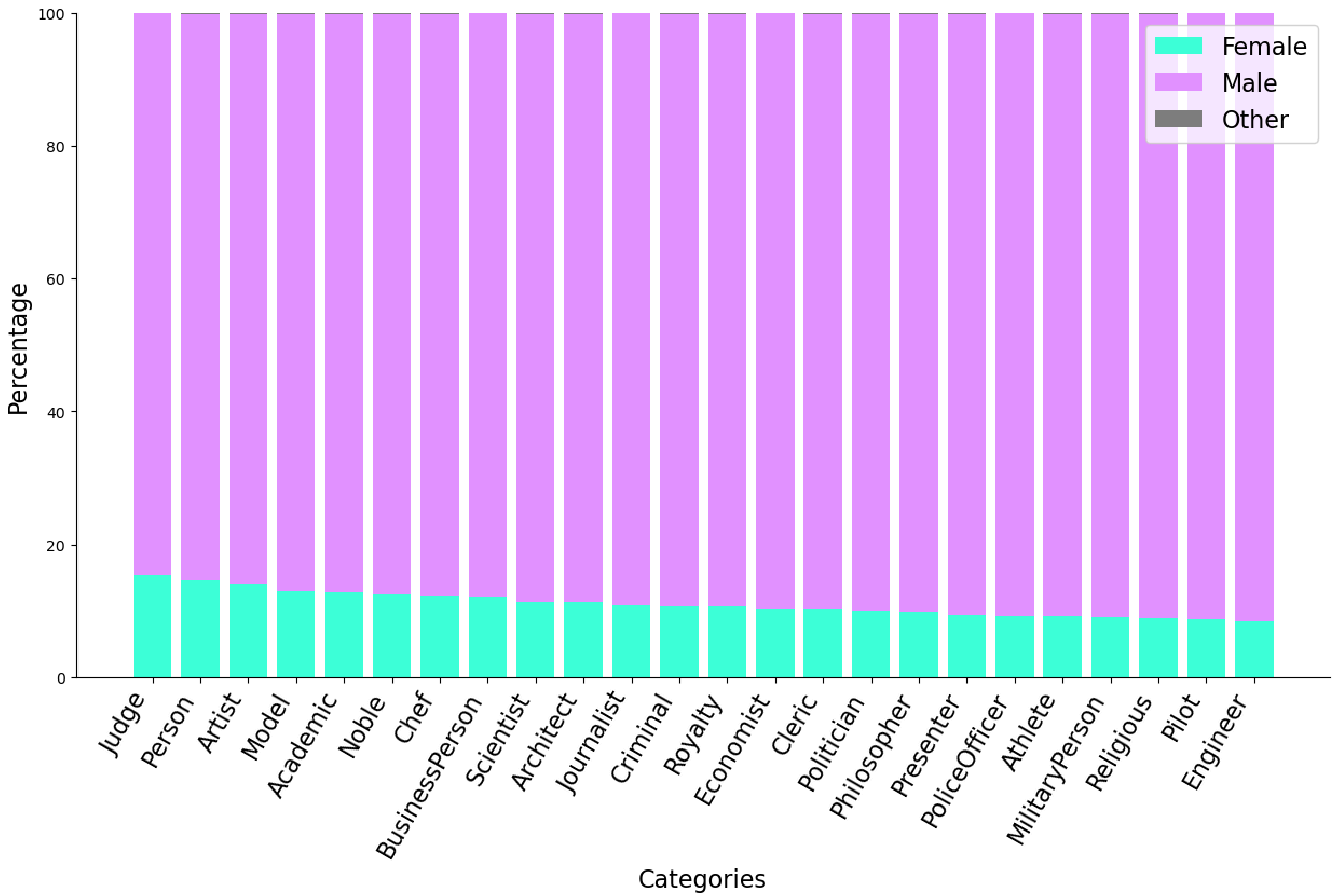
Observations:

- Male: 85.45%
- Female: 14.42%
- Non-binary: 0.13%

Limitations:

- Not enough data in "other" gender --> so only binary study.
- Unbalanced data.

Gender distribution of editors by category



BACK TO OVERVIEW →

Methods and results: What

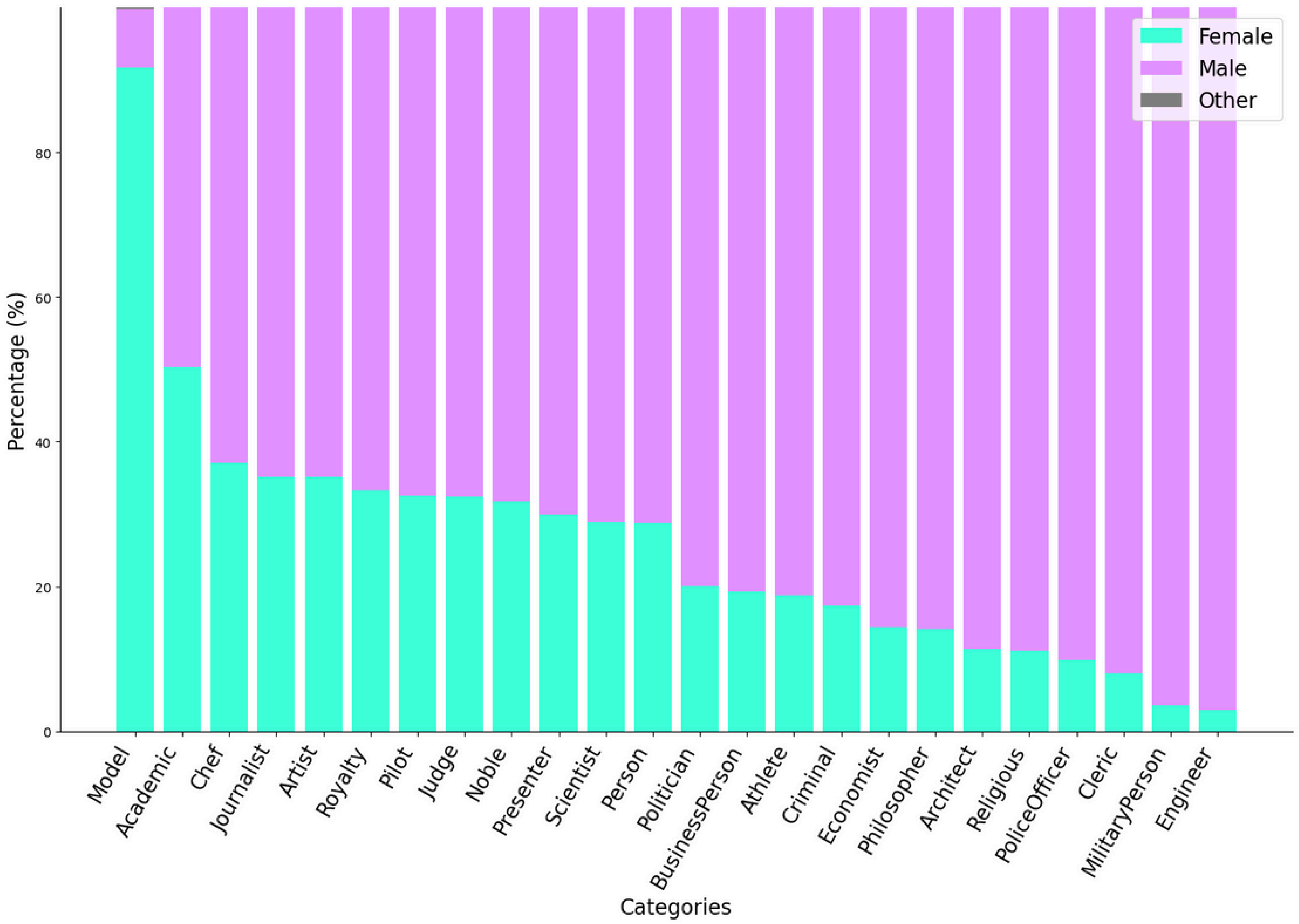
Observations:

- Male: 74.63%
- Female: 25.35%
- Non-binary: 0.02%

Limitations:

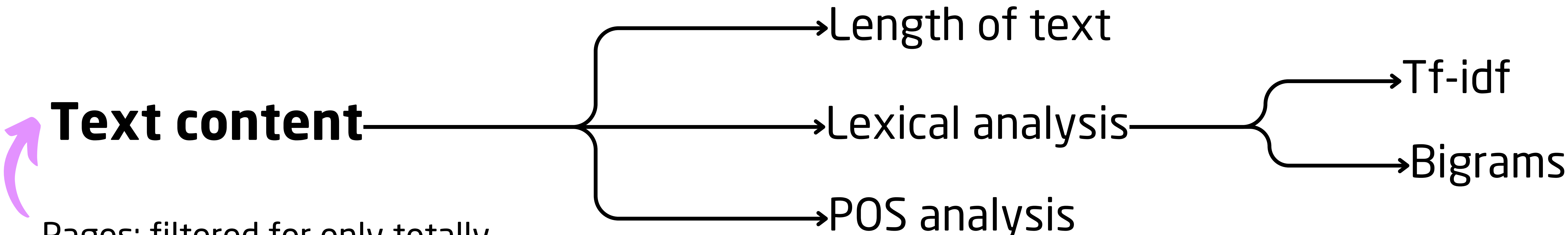
- Unbalanced data in some categories.
- Not enough "Other" gender.

Gender distribution of biographies by category



BACK TO OVERVIEW →

Methods and results: How



Text content

Length of text

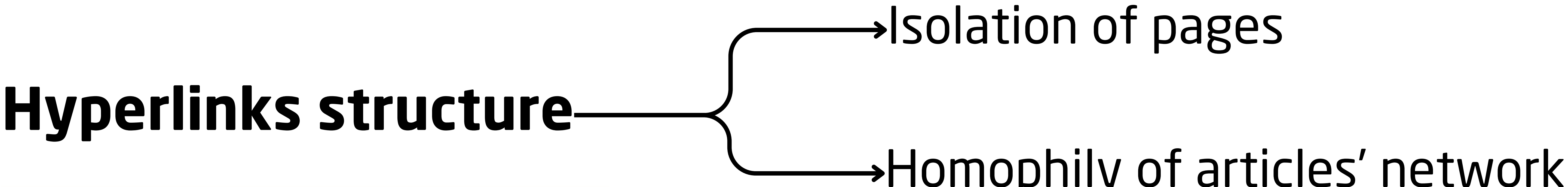
Lexical analysis

Tf-idf

Bigrams

POS analysis

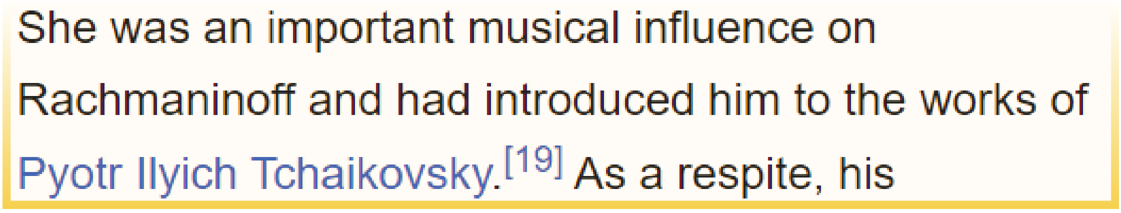
Pages: filtered for only totally-edited-by-one-gender pages.



Hyperlinks structure

Isolation of pages

Homophily of articles’ network



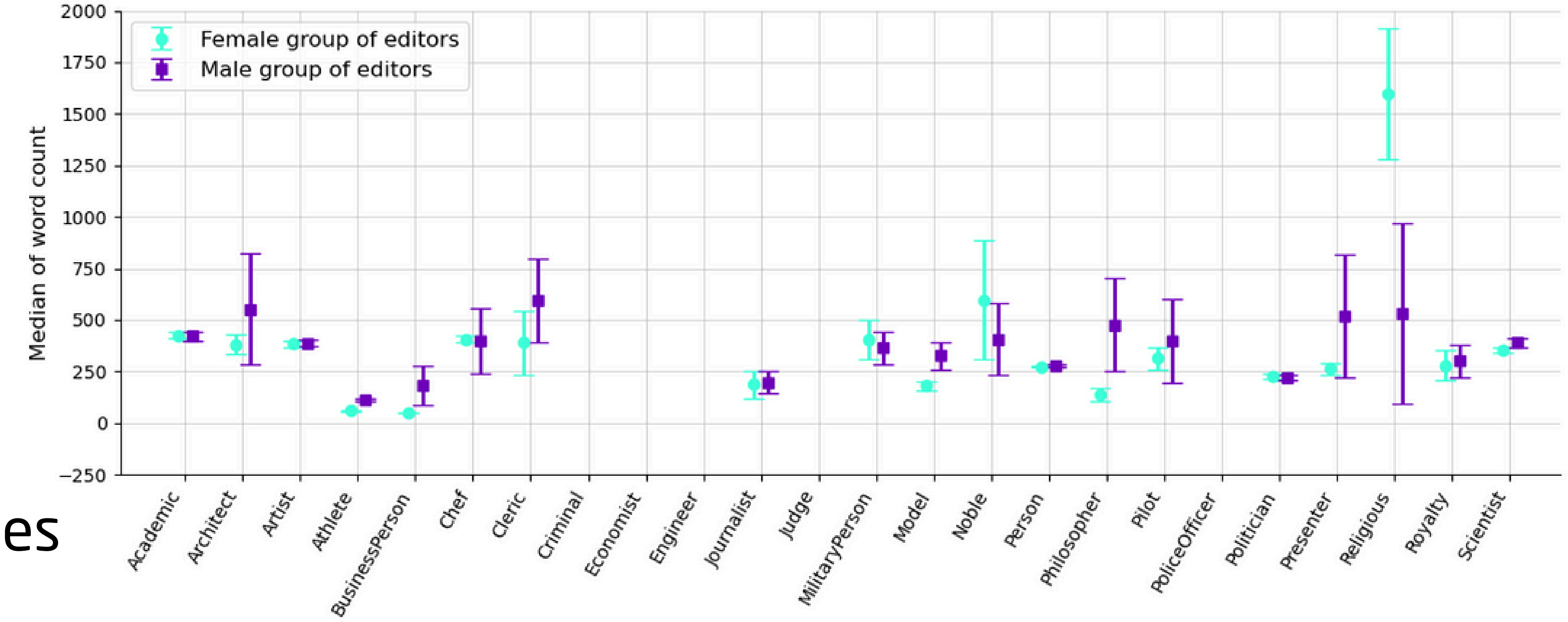
She was an important musical influence on Rachmaninoff and had introduced him to the works of Pyotr Ilyich Tchaikovsky.[19] As a respite, his

BACK TO OVERVIEW →

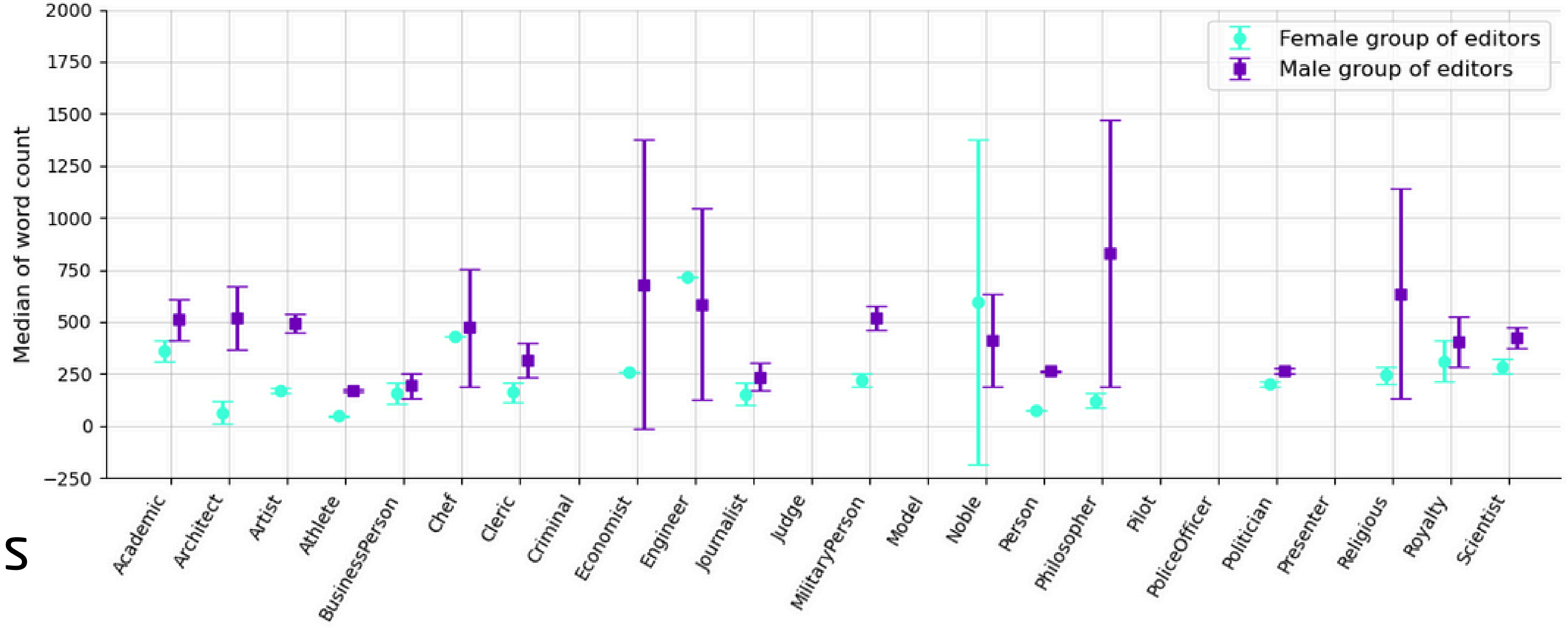
Length of text

Median word count

95% CI



Female pages



Male pages

Motivation:

- Quantity of information.

Observations:

- No difference between female and male biographies.
- In male pages: female contributors write significantly fewer words than their male counterparts.

Limitations:

- How many editors
- Creation or editing of pages.

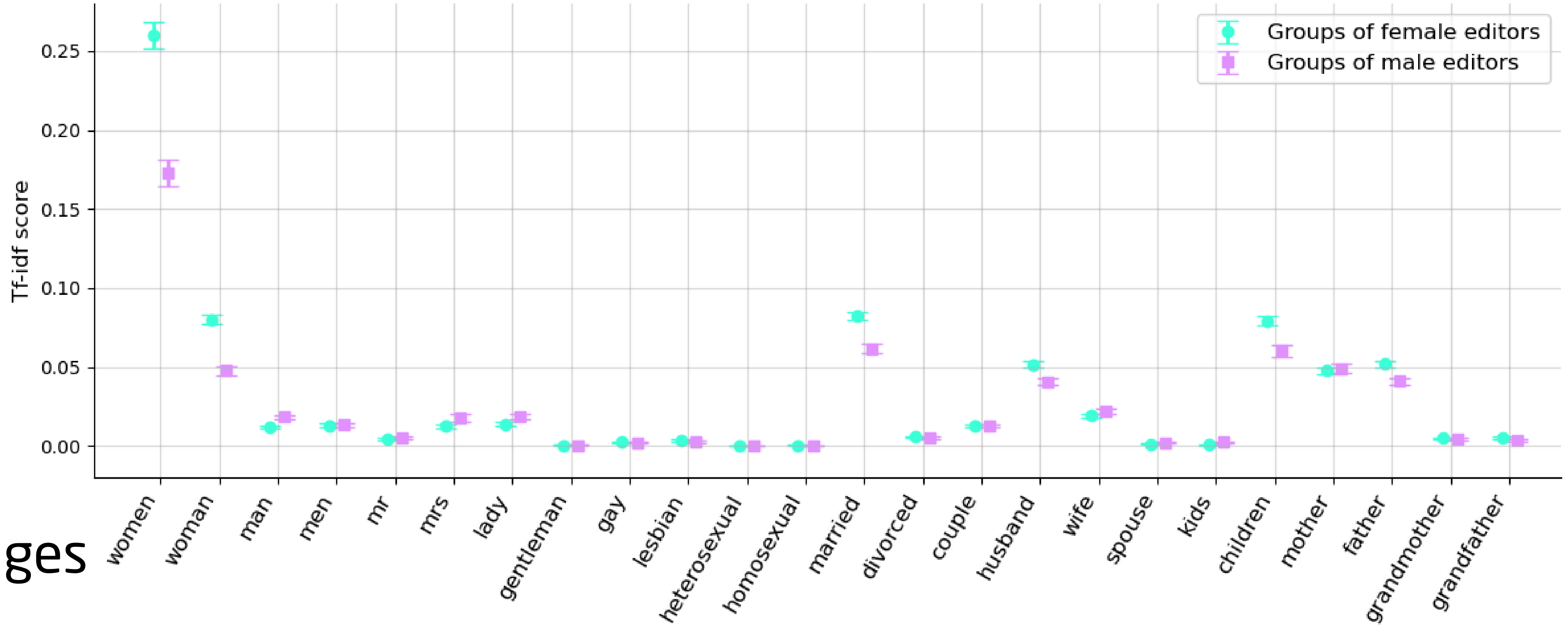
BACK TO OVERVIEW →

Tf-idf scores

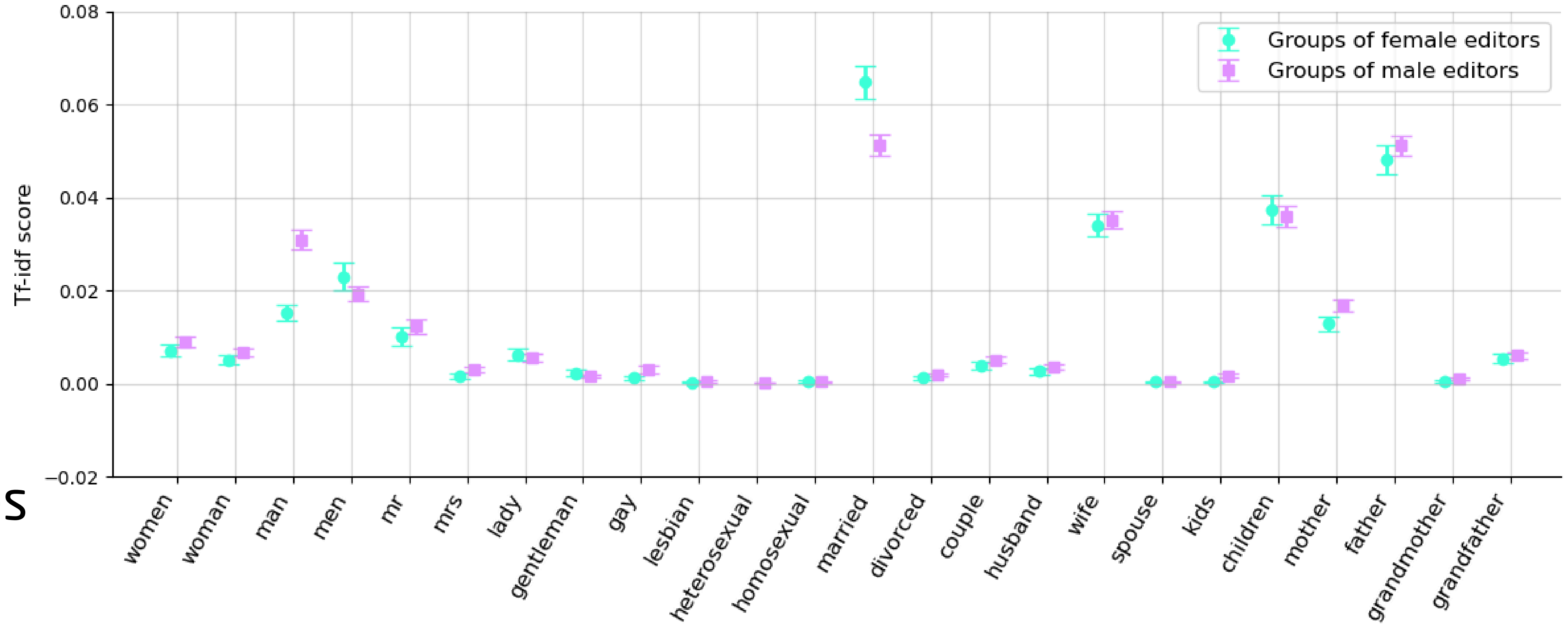
Tf-idf score 95% CI for selected words in Person category

Motivation:

- Relevance of words.



Female pages



Male pages

Observations:

- Female contributors tend to give gender and famiy related words more importance than male in women's biographies.
- The term *woman* in female articles and *man* in male articles reveals that *woman* holds significantly greater relevance.

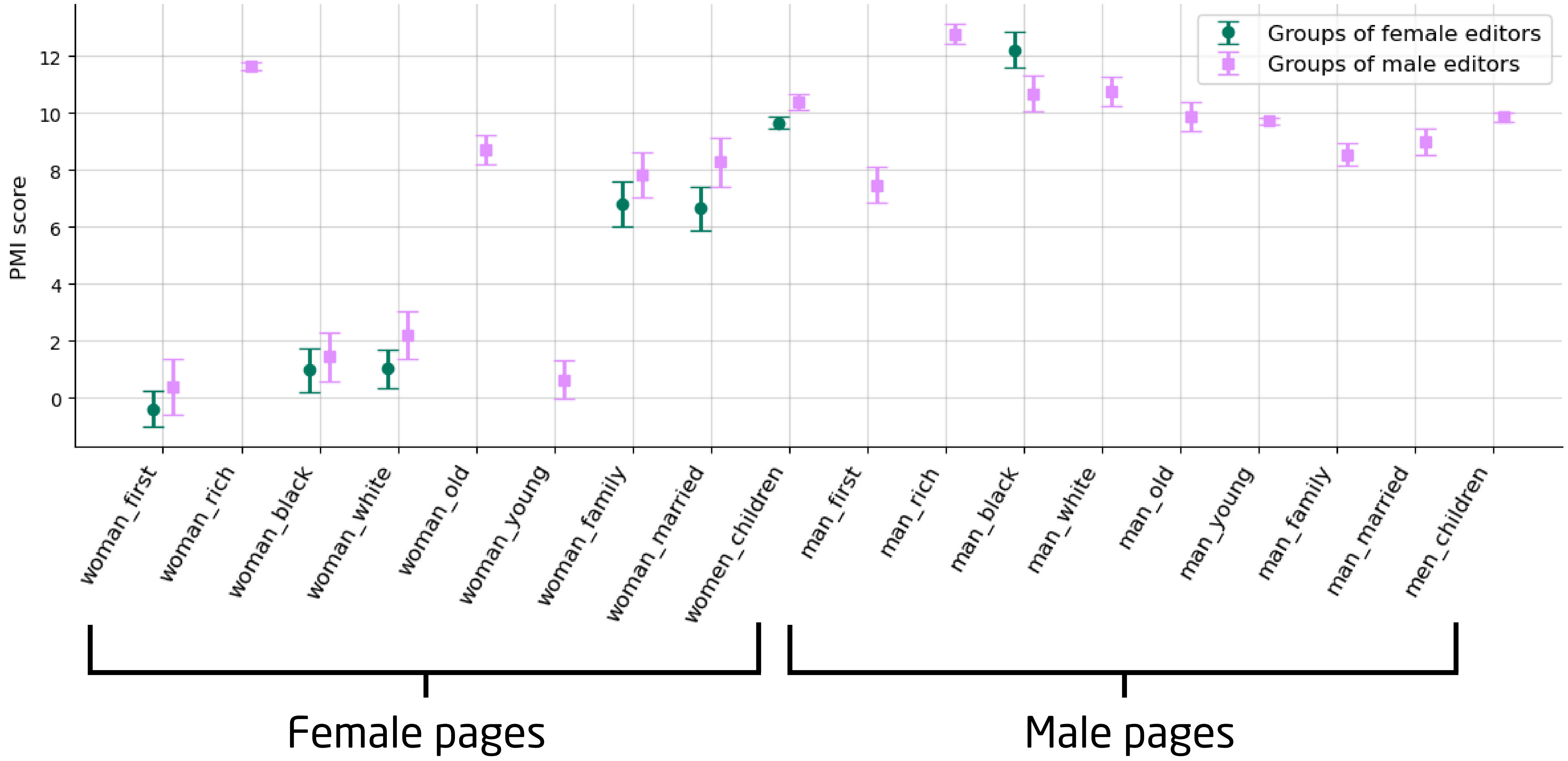
Limitations:

- Pre-selected group of words

BACK TO OVERVIEW →

Bigrams of words

PMI score 95% CI in Person category



Motivation:

- Context of words.

Observations:

- Male editors more frequently associate the word *woman* with family terms compared to female editors.

Limitations:

- Pre-selected group of words
- PMI has a tendency to give higher scores to low-frequency events

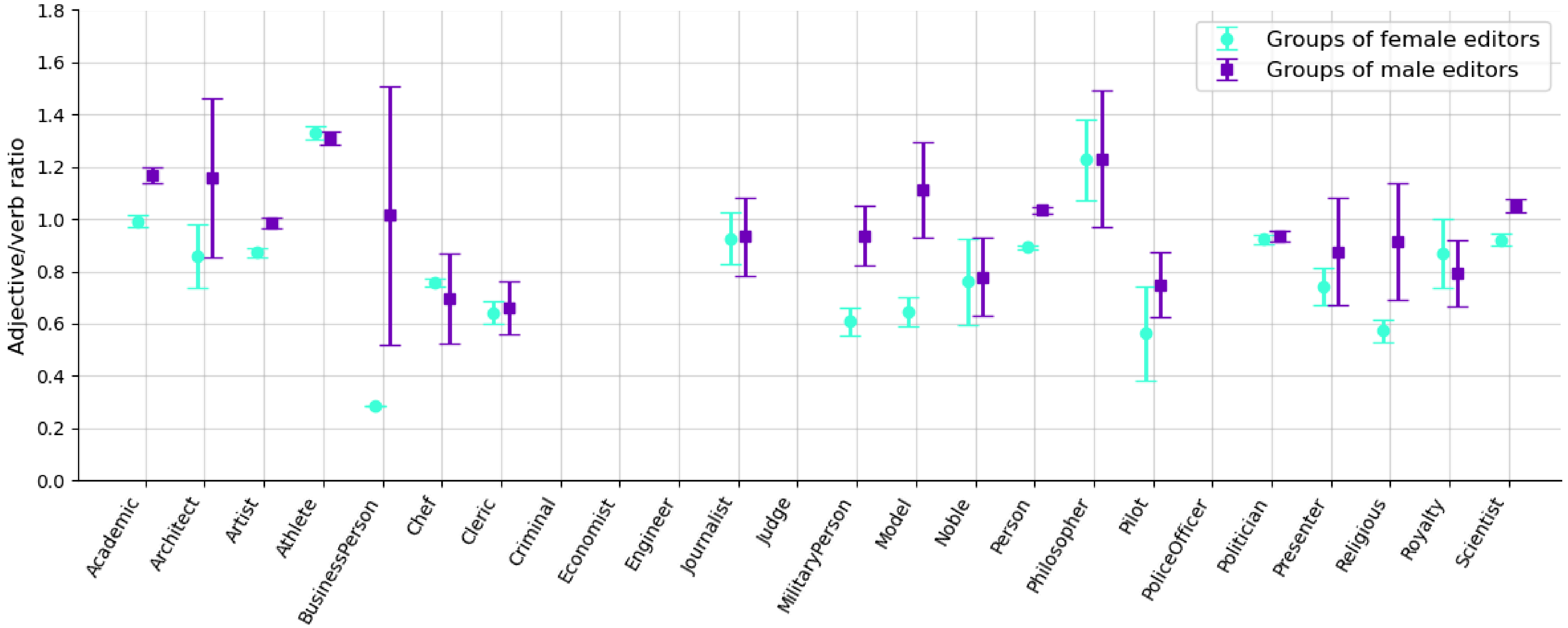
Further:

- n-grams with n>2

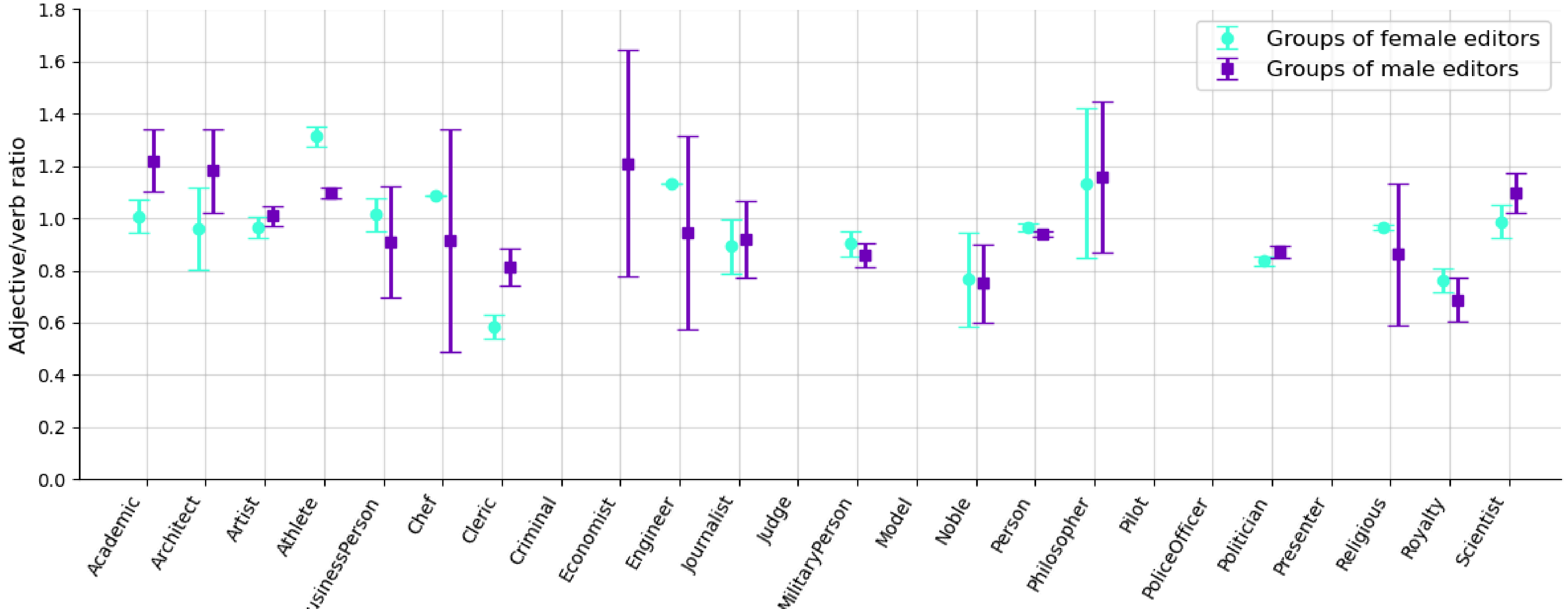
BACK TO OVERVIEW →

POS analysis

Adjective-to-verb ratio 95% CI



Female pages



Male pages

Motivation:

- Abstract or concrete language?

Observations:

- Female biographies: the ratio is significantly higher in groups of male editors compared to female editors.
- Male biographies: the difference between these two groups of editors is minimal.

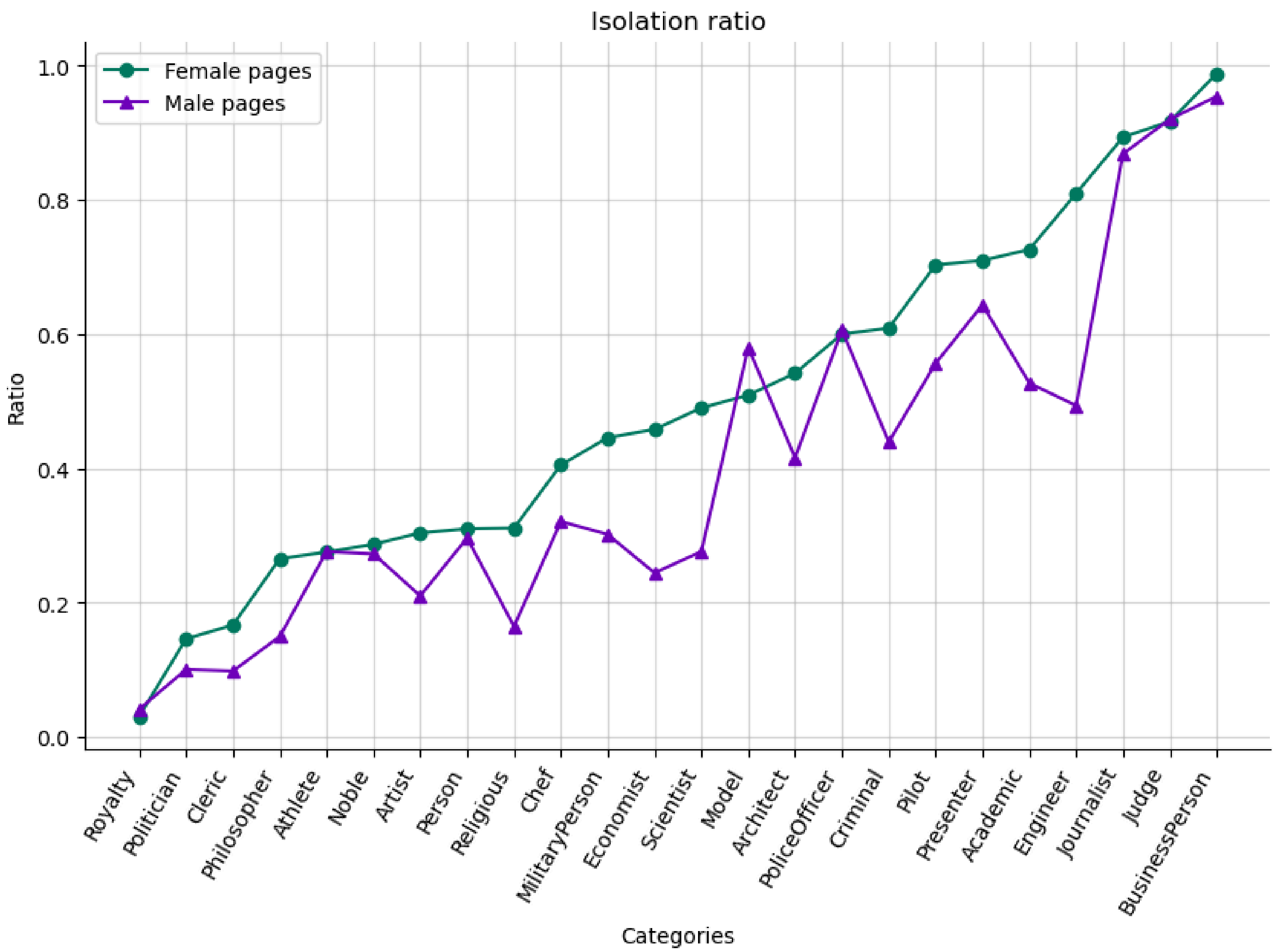
Further:

- Other POS-tags reflect other types of language usage

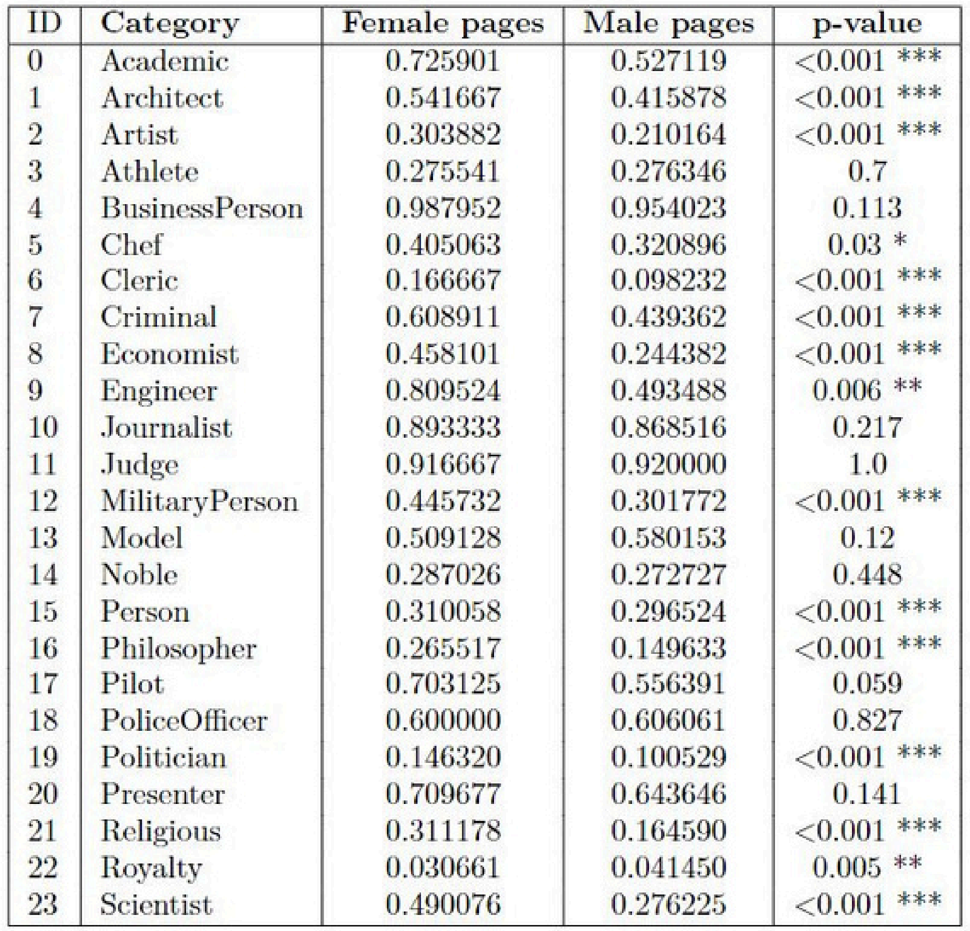
BACK TO OVERVIEW →

Isolation of pages

Isolation ratio per category and per biography gender



Permutation test results



ID	Category	Female pages	Male pages	p-value
0	Academic	0.725901	0.527119	<0.001 ***
1	Architect	0.541667	0.415878	<0.001 ***
2	Artist	0.303882	0.210164	<0.001 ***
3	Athlete	0.275541	0.276346	0.7
4	BusinessPerson	0.987952	0.954023	0.113
5	Chef	0.405063	0.320896	0.03 *
6	Cleric	0.166667	0.098232	<0.001 ***
7	Criminal	0.608911	0.439362	<0.001 ***
8	Economist	0.458101	0.244382	<0.001 ***
9	Engineer	0.809524	0.493488	0.006 **
10	Journalist	0.893333	0.868516	0.217
11	Judge	0.916667	0.920000	1.0
12	MilitaryPerson	0.445732	0.301772	<0.001 ***
13	Model	0.509128	0.580153	0.12
14	Noble	0.287026	0.272727	0.448
15	Person	0.310058	0.296524	<0.001 ***
16	Philosopher	0.265517	0.149633	<0.001 ***
17	Pilot	0.703125	0.556391	0.059
18	PoliceOfficer	0.600000	0.606061	0.827
19	Politician	0.146320	0.100529	<0.001 ***
20	Presenter	0.709677	0.643646	0.141
21	Religious	0.311178	0.164590	<0.001 ***
22	Royalty	0.030661	0.041450	0.005 **
23	Scientist	0.490076	0.276225	<0.001 ***

Table C.5: Isolation ratio for the pages in every category. The last column contains the p-values. Statistical comparisons were performed using a permutation test where null hypothesis is that the isolation ratio is the same under both genders of pages (*** $p < 0.001$, ** $p < 0.01$, * $p<0.05$).

Motivation:

- Retrieval and relevance of pages.

Observations:

- Women's articles are consistently more isolated than men's.
- Athlete category: surprising exception.

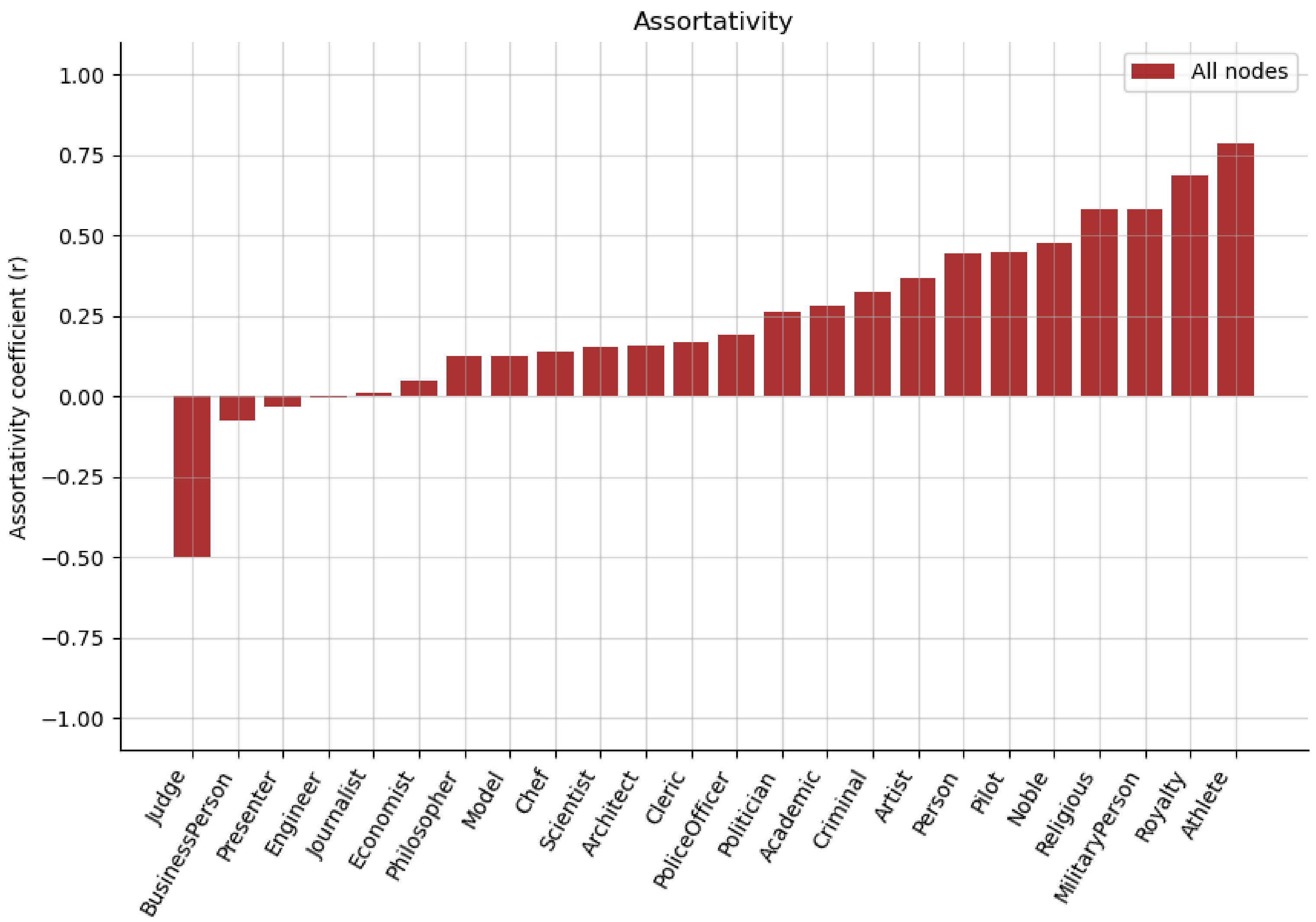
Further:

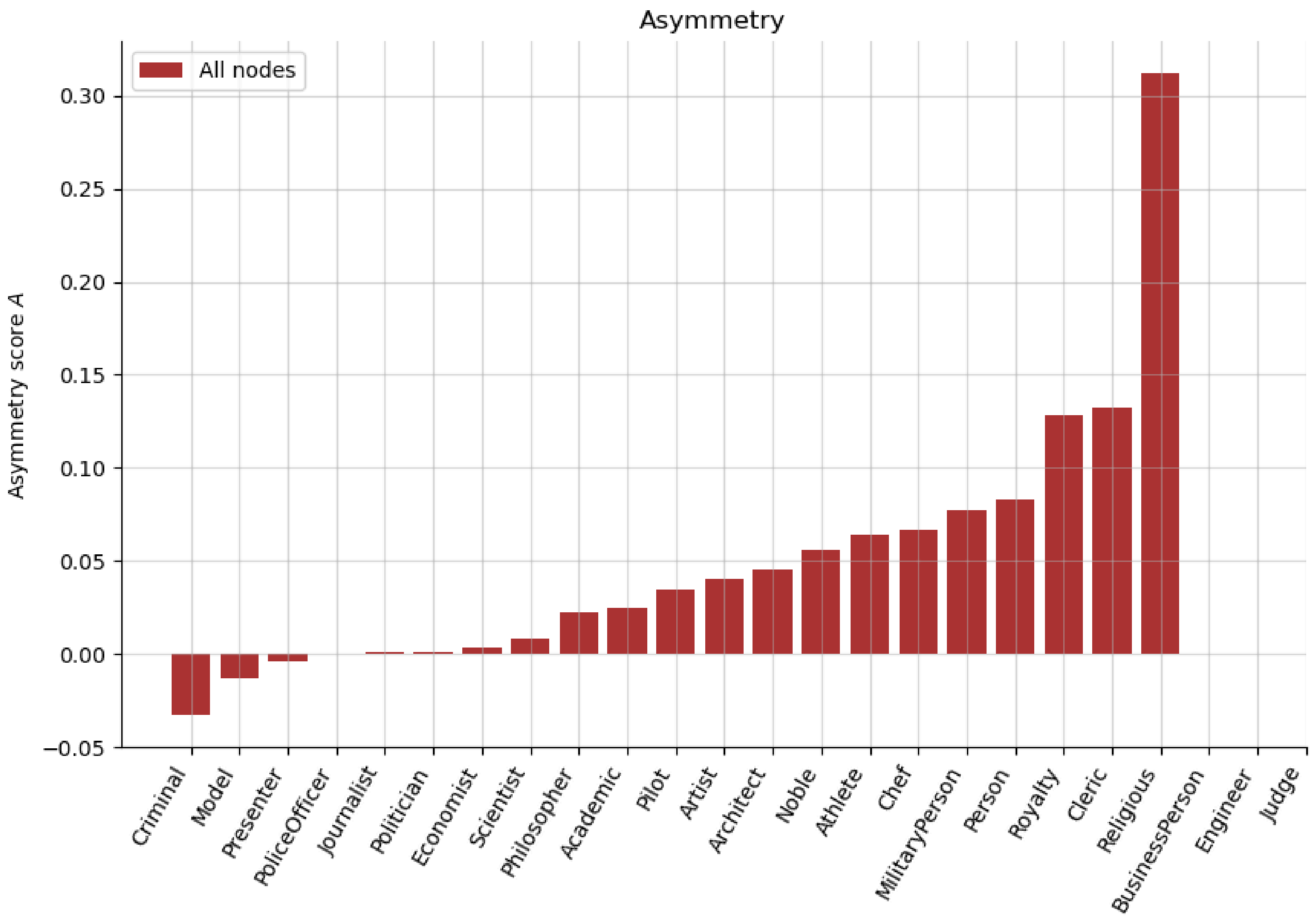
- For directed network: analyse in-degree and out-degree separately.

BACK TO OVERVIEW →

Homophily

Assortativity and asymmetry of the network per category





Motivation:

- Retrieval and relevance of pages.

Observations:

- Nodes tend to connect to other nodes that are similar in the gender attribute.
- Biographies women are more likely to show a link to articles about male personalities than the other way around.

Further:

- Analyse also editors trends.
- Structural biases can also manifest in the centrality measures.

BACK TO OVERVIEW ⟶

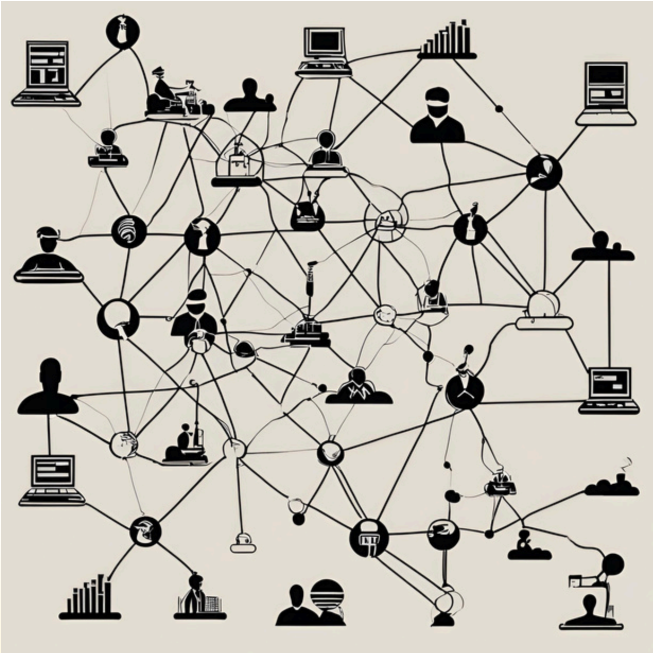
SUMMARY AND CONCLUSION

Unbalanced representation of genders in editors and pages.

Specially in the editors community in Wikipedia

Language bias

Diferences in lengths of texts, lexical analysis and POS-analysis, indicating differences in how editors portray various personalities.



Gender imbalance in connectivity of articles.

Women pages tend to be much less connected than men's.

BACK TO OVERVIEW →

FUTURE RESEARCH

LITTLE CHANGES...

- Word count related to number of editors.
- Change PMI measure for Bigrams of words.
- Test homophily results to know how significant they are.

- Different language versions of Wikipedia.
- Separate revisions and analyse one by one about the editor and what they did.
- Historical change: separate pages into different born periods.
- Other human dimensions (age, nationality, socio-economic status, etc).
- Expand to not only Wikipedia, but in general.

BACK TO OVERVIEW →

REFERENCES

- Claudia Wagner et al. “It’s a man’s Wikipedia? Assessing gender inequality in an online encyclopedia.” In: Proceedings of the international AAAI conference on web and social media. Volume 9. 1. 2015, pages 454–463.
- Claudia Wagner et al. “Women through the glass ceiling: gender asymmetries in Wikipedia.” In: EPJ data science 5 (2016), pages 1–24.
- Anne Maass et al. “Language use in intergroup contexts: The linguistic intergroup bias.” In: Journal of personality and social psychology 57.6 (1989), page 981.
- Pablo Beytía et al. “Visual gender biases in wikipedia: A systematic evaluation across the ten most spoken languages.” In: Proceedings of the International AAAI Conference on Web and Social Media. Volume 16. 2022, pages 43–54.
- Heather Ford and Judy Wajcman. “‘Anyone can edit’, not everyone does: Wikipedia’s infrastructure and the gender gap.” In: Social studies of science 47.4 (2017), pages 511–527.

Q&A SESSION

Thank you for listening!

BACK TO OVERVIEW →